图书情报工作
第 62 卷 第 5 期 2018 年 3 月


# 基于多属性加权的社会化问答社区关键词提取方法*

余本功[1,2] 李婷[1] 杨颖[1]
[1] 合肥工业大学管理学院 合肥 230009 [2] 合肥工业大学过程优化与智能决策教育部重点实验室 合肥 230009

**摘要：**［目的/意义］现有的关键词提取方法不适应社会化问答社区文本长度较短、内容表述口语化、数据集稀疏的特点，且很少考虑用户关注程度对词语重要性的影响，不能有效地提取此类文本的关键词，因此，提出针对社会化问答社区的多属性加权关键词提取方法。［方法/过程］多属性加权关键词提取方法通过引入调节函数和词性对传统 TF-IDF 进行改进，并通过线性加权融合用户回答数、关注数、浏览数以及评论数 4 个用户关注属性来综合度量词语权重。［结果/结论］实验表明，该方法能更有效地提取社会化问答社区文本的关键词。

**关键词：**社会化问答社区 关键词提取 TF-IDF 多属性加权
**分类号：**TP391
**DOI**：10.13266/j.issn.0252－3116.2018.05.015



## 1 引言

随着信息技术的发展和互联网的全面普及，以知乎、Quora 等为代表的社会化问答社区成为人们信息交流和知识共享的重要渠道[1]。社会化问答社区是传统问答网站和虚拟社区结合的产物，支持用户围绕共同兴趣和目标自我生成内容，用户既是信息资源的受益者，又是信息资源的建设者[2]。社会化问答社区中的文本充分反映了用户的知识面、兴趣爱好等信息，对网络舆情分析、用户兴趣挖掘、社区知识发现等自然语言处理相关研究有重要价值。关键词提取是自然语言处理的基础和核心之一，对自然语言处理技术的应用效果有重要影响。

目前，主流的关键词提取方法有三类：基于机器学习的方法[3－4]、基于语义的方法[5]以及基于统计的方法[6]。基于机器学习的方法通过训练机器学习模型自动提取关键词，是建立在大量的语料库基础上的，需要大量的参数训练来保证结果的准确性；基于语义的方法通过构建词语间的语义关系网络来分析和提取关键词[7]，由于缺少语义定义标准，该方法易受主观性影响，且对背景知识库、词典和词表依赖较高，对文本格式有严格要求。由于用户生成内容的自由性，社会化问答社区的文本长度较短、内容表述口语化、数据集稀疏[8]，且文本更新速度快[9]，很难建立标准的语料库和背景知识库。因此，这两种方法不适用于社会化问答社区文本的关键词提取。

基于统计的方法通过统计文本特征来提取关键词[10]，其中应用最多的是 TF-IDF（Term Frequency-inverse document frequency）方法。该方法简单通用，对文本长度和语言规范限制较少，但准确性不高[11－12]。针对这个问题，学者们进行了大量探索。研究成果显示，在词频分析的基础上融入词性[13]、词语关联度[11,14]、词语位置[11,15]、词跨度[15]等属性，能有效避免传统关键词提取方法产生的误差。此外，在基于虚拟社区的研究中，有学者发现，用户是信息的生产者、传播者和使用者，用户浏览、回复等数据记录体现了用户对该内容的关注程度，衡量词语重要性的时候应将这些属性纳入考虑范围[12,16]。目前，引入属性提高关键词提取效率的研究取得了一定成果，但现有方法很少考虑其对社会化问答社区文本的适用性，不能有效应用于该文本集的处理。一方面，现有方法从中文文本或网页文本出发，不一定适用社会化问答社区文本长度较短、表述口语化、数据集稀疏的特点；另一方面，与其他来源的数据集相比，社会化问答社区文本的结构不同，用户关注度体现的形式不同，衡量词语重要性的

＊ 本文系国家自然科学基金项目“基于制造大数据的产品研发知识集成与服务机制研究”（项目编号：71671057）和“不确定环境下的复杂产品研发协同绩效动态评价研究”（项目编号：71573071）研究成果之一。
作者简介：余本功（ORCID：0000－0003－4170－2335），教授，博士，E-mail：bgyu@hfut.edu.cn；李婷（ORCID：0000－0002－5556－7624），硕士研究生；杨颖（ORCID：0000－0002－9912－3443），副教授，博士。
收稿日期：2017－08－23 修回日期：2017－11－14 本文起止页码：132－139 本文责任编辑：徐健

具体属性存在区别。因此,如何综合度量社会化问答社区文本的属性,提出适用于社会化问答社区实际的关键词提取方法是尚待研究的重要问题。

## 2 相关关键词提取方法

### 2.1 基于词频的关键词提取方法

TF-IDF[17-18] 是词频权重计算使用最多的方法。TF 表示词语出现的次数,IDF 表示含有某词语的文本占文本集的比例。给定文本集 $P=\{(p_j)\mid j=1,2,\cdots,N\}$,记文本集中所有词语构成的集合为 $T=\{(t_i)\mid i=1,2,3,\cdots\}$,词语 $t_i$ 的权重计算公式[16]为:

$$W_{ij}=tf_{ij}\times idf_i=tf_{ij}\times\log\frac{N}{n}\tag{1}$$

其中,$n$ 表示文本集 $P$ 中含有词语 $t_i$ 的文本个数,$tf_{ij}$ 表示词语 $t_i$ 在文本 $p_j$ 中出现的次数。

TF-IDF 依据的原理为:词语出现的次数越多越重要;当一个词语在某一个文本中多次出现,而在其他文本中很少出现,说明这个词能很好地区别该文档,在文本中的重要性越高。虽然这种方法简单有效,但是不能在各种场合取得好的应用结果[14,19]。因此,很多学者结合具体情况深入分析后,对 TF-IDF 进行了改进。罗燕等提出结合同频词数统计规律的改进 TF-IDF 关键提取方法[11];张建娥等将 TF-IDF 和词语关联度结合,用于中文文本关键词提取[14];张瑾等引入词位置和词跨度对 TF-IDF 方法进行改进,用于情报关键词提取[15];罗繁明等在情报关键词提取中构建了综合词偏度、词语位置权重和 TF-IDF 的关键词重要性评估函数[20];钱爱兵等在新闻网页关键词提取中,将 TF-IDF 和词长、位置等因素进行加权,得到词语的综合排序[21];张保富考虑到词语在类间和类内的分布情况,采用特征项在类间和类内信息分布熵来调整 TF-IDF 的权重计算[22]。

### 2.2 引入词性属性的关键词提取方法

一般地,动词、名词、形容词、副词能够表示文本的主要信息,助词、连词、代词等虚词主要用于修饰语句,对概括文本信息没有很大价值。目前,很多学者就词性属性对关键词提取效果的影响进行了研究。张建娥等根据人工标注结果对关键词的词性进行了统计,发现名词、动词、形容词、副词四类词性的关键词数量和达到关键词总数的 95.5%,并在此基础上提出词频、词语关联性、词性和位置特征线性加权的多特征融合关键词提取方法[14];袁津生等提出综合中文新闻网页的统计特征、位置特征和词性特征等在内的多特征综合关键词权重计算方法,该方法对名词、动词、时间词、方位词、形容词、副词赋予不同的权重[23];蒋昌金等构建了词频、词性、词的位置、词长等因素的加权计算公式用于提高能够表达主题的词语的权重,其中名词和名词词组被赋予较高的权重[24];李湘东等结合词性、位置属性对词语权重进行修正并应用到 LDA 生成模型中,用于抽取文本的粗粒度特征[25];路永和等提出受词性影响的特征权重计算方法[26];周鹏在微博舆情研究中提出增加中心度、词性、词位置属性的关键词抽取方法[27]。

### 2.3 引入用户关注属性的关键词提取方法

虚拟社区中,用户的浏览、评论等行为是自由的,用户的这些行为以浏览数、评论数等数据的形式被记录。用户的兴趣和关注点不同,导致不同文本的用户行为数据有较大差别。因此,与来自文献的文本不同,虚拟社区中的文本除了具有词频、词性等词语本身的属性之外,还有用户浏览数、回复数等用户关注属性。目前,有学者提出虚拟社区中词语的重要性不仅仅取决于词语出现的频率,还取决于其受用户关注的程度。黄鲁成等在社会化问答社区话题识别研究中引入用户对问题的关注数和回答数来衡量词语重要性,并统计观察用户关注数和回答数的数字规律对用户关注情况进行量化[8]。廖晓等认为企业虚拟社区中词语的重要性受到词频、用户浏览数和回复数的影响,并结合媒体关注度计算方法对用户浏览数和回复数进行计算,从词频和用户关注度两个方面综合分析词语重要性[16]。

社会化问答社区是兼具问答和社交功能的平台,对用户全面开放提问、回答、最佳答案选择等过程。在社会化问答社区中,用户通过回答功能分享知识,通过关注功能实时了解话题变化,通过评论功能表达对提问内容的看法。因此,用户对文本的关注情况体现在用户浏览数、用户关注数、用户回答数、用户评论数 4 个方面。同时,由于用户在社会化问答社区发表言论的自由性,产生的数据呈现出文本长度较短、文本表述口语化、数据集稀疏的特点。目前,基于统计的关键词提取研究取得了一定成果,但由于考虑到的属性不足或不适应社会化问答社区文本的特点,现有关键词提取方法不能有效应用于该类文本集的关键词提取。

## 3 社会化问答社区多属性加权关键词提取方法

### 3.1 关键词提取方法和流程

针对现有关键词提取方法应用于社会化问答社区

的不足，本文提出结合社会化问答社区特征的多属性加权关键词提取方法（Keywords Extraction Method based on Multi-attributes Weighted，MW-KEM），基本流程见图 1。词语出现频次的大小能在一定程度上反映词语的重要性；通过给不同词性的词语赋予不同的权重有助于凸显有效词语，能有效提高关键词提取效率；用户关注度越高的词语越能反映文本的内容，而社会化问答社区中，用户的关注程度体现在回答数、关注数、浏览数、评论数 4 个方面。因此，本方法以词频（FR）、词性（POS）、回答数（RE）、关注数（AT）、浏览数（BR）、评论数（CO）6 个属性为指标，通过线性加权综合度量词语重要性。提出词语 $t_i$ 在文本 $p_j$ 中的词语权重 $W_{ij}$ 计算公式：

$$W_{ij}=\partial_1\times FP_{ij}+\partial_2\times RE(p_j)+\partial_3\times AT(p_j)+\partial_4\times BR(p_j)+\partial_5\times CO(p_j) \quad (2)$$

词语 $t_i$ 在文本集 $P$ 中的权重为：

$$W_i=\sum_{j=1}^{N}W_{ij} \quad (3)$$

其中，$FP_{ij}$ 为词频和词性的综合权重；$RE(p_j)$、$AT(p_j)$、$BR(p_j)$ 和 $CO(p_j)$ 为用户关注属性权重，使用 TF-PDF（Term Frequency - Proportional Document Frequency）[28] 话题关注度计算方法量化；$\partial_m$，$m=1,2\cdots,5$ 为各属性的权重系数，且 $\sum_{m=1}^{5}\partial_m=1$，使用层次分析法（Analytic Hierarchy Process，AHP）[29-30] 确定。

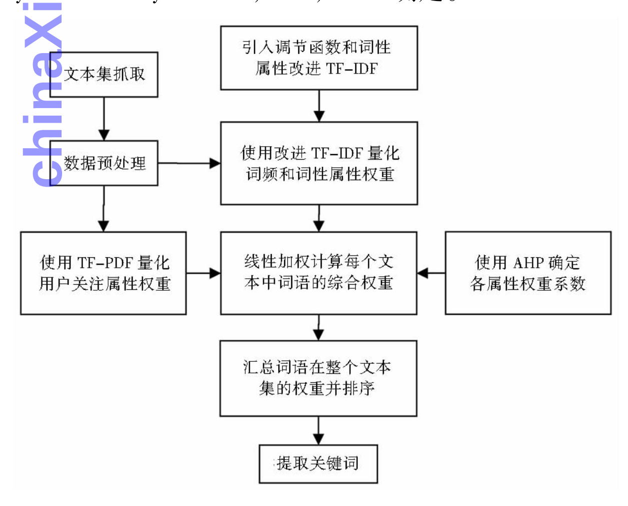



图 1 多属性加权关键词提取方法流程

## 3.2 基于改进 TF-IDF 的频率与词性权重量化

由于社会化问答社区文本长度较短，TF 值较小，传统 TF-IDF 的计算结果受 IDF 的影响较大，容易出现两方面的不足：①文本集 $P$ 中含有词语 $t_i$ 的文本个数 $n$ 接近文本集 $P$ 的文本总数 $N$ 时 IDF 值很低，整个权重值过小，导致一些虽然在多篇文章中出现，但能很好地表达文本特征的词语不能被选为关键词；②$n$ 接近 0 时 IDF 值很高，整个权重值偏大，导致一些低频词被误选为关键词[12]。为了解决这个问题，本文基于幂函数 $y=x^3$ 对 $n$ 值进行调节，提高 TF-IDF 在 $n$ 值较大时的计算结果，降低 TF-IDF 在 $n$ 值较小时的计算结果。令 $n'=a(n-N/2)^3+b$，其中 $a$、$b$ 为常量。为了避免取值范围的变化影响有效性，令函数端点为 $(0,0)$、$(N,N)$，得到 $a$、$b$ 值分别为 $4/N^2$、$N/2$。故调节函数为：

$$n'=(2/N)^2(n-N/2)^3+N/2 \quad (4)$$

图 2、图 3 分别以 $N$ 取 1 000 为例对调节函数和调节后的 IDF 函数进行展示。从图 3 中可以看出，当 $n<N/2$ 时，IDF 值比传统 IDF 方法计算的值小，$n$ 取值较小时变化比较明显；当 $n>N/2$ 时，IDF 值比传统 IDF 方法计算的值大，$n$ 取值较大时变化比较明显。

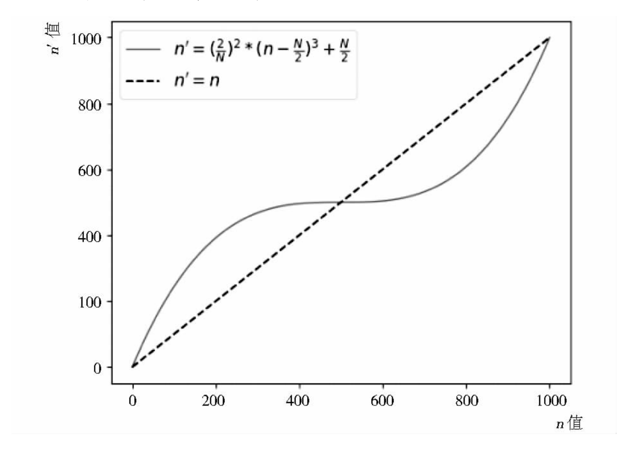



图 2 引入的调节函数

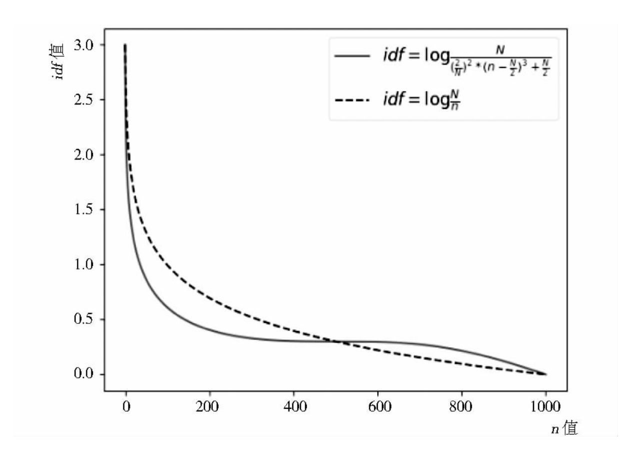



图 3 调节后的 IDF 对比

为进一步提高关键词提取能力，根据词性属性给词语赋予相应的权重。通常，名词、动词、形容词、副词四种词性的词语占关键词集合的绝大多数，而虚词、连词、助词等词性的词语主要用于加强语句，不能起到总结文本内容的作用。因此，本方法对名词、动词、形容词、副词四种词性的词语赋予较高的权重，对其他词性

的词语权重赋值为 0。同时,将词性属性融合到 TF-IDF 计算过程中,得到改进后的 TF-IDF 如下:

$$fp_{ij} = pos \times tf_{ij} \times idf_i = pos \times tf_{ij} \times \log N/n' \quad (5)$$

其中,$n$、$tf_{ij}$与传统 TF-IDF 含义相同,$fp_{ij}$表示词语 $t_i$ 在文本 $p_j$ 中词频与词性的综合权重。$pos$ 表示不同词性词语的权重,这里参照同类方法的常用取值,赋动词、名词为 1.5,形容词、副词为 1。为了便于各个属性之间的比较,取文本 $p_j$ 中 $fp_{ij}$的最大值和最小值,对词频和词性的综合权重做标准化处理:

$$FP_{ij} = \frac{fp_{ij} - \min(fp_{ij})}{\max(fp_{ij}) - \min(fp_{ij})} \quad (6)$$

## 3.3 基于 TF-PDF 的用户关注属性权重量化

社会化问答社区中,用户关注程度主要体现在用户对问题的回答数、关注数、浏览数、评论数上。此处引入话题关注度计算方法 TF-PDF[28] 对这 4 个属性进行量化:

$$\omega_k(p_j) = \frac{a_k(p_j)}{\sqrt{\sum_{j=1}^{N}(a_k(p_j))^2}} \cdot \exp\left(\frac{a_k(p_j)}{\sum_{j=1}^{N} a_k(p_j)}\right),\ k = 1,2,3,4 \quad (7)$$

其中,$a_k(p_j)$,$k=1,2,3,4$ 分别对应文本 $p_j$ 的回答数、关注数、浏览数以及评论数;$\omega_k(p_j)$,$k=1,2,3,4$ 分别对应文本 $p_j$ 的回答属性权重 $RE(p_j)$、关注属性权重 $AT(p_j)$、浏览属性权重 $BR(p_j)$ 以及评论属性权重 $CO(p_j)$。

## 3.4 基于 AHP 的属性权重系数赋值

采用层次分析法确定各属性权重的系数。层次分析法是美国匹兹堡大学教授 T. L. Saaty 提出的确定指标权重的常用有效方法,一般由 4 个步骤构成:建立层次结构模型、构造判断矩阵、层次单排序及层次总排序[29－30]。

3.4.1 建立层次结构模型 构建层次分析法结构模型见图 4。目标层为给词语赋予合适的权重;准则层由词语属性和用户关注属性两个大类构成;方案层为待确定权重的各个具体属性。

3.4.2 构造判断矩阵 采用一致矩阵法构建判断矩阵,矩阵元素采用 1－9 标度。邀请 5 位专家对各层指标的重要性打分,综合各专家意见得到判断矩阵如表 1、2 所示:

表 1 A 判断矩阵

| | $B_1$ | $B_2$ | W | CR |
|---|---|---|---|---|
| $B_1$ | 1 | 1 | 0.5 | 0 <0.1 |
| $B_2$ | 1 | 1 | 0.5 | |

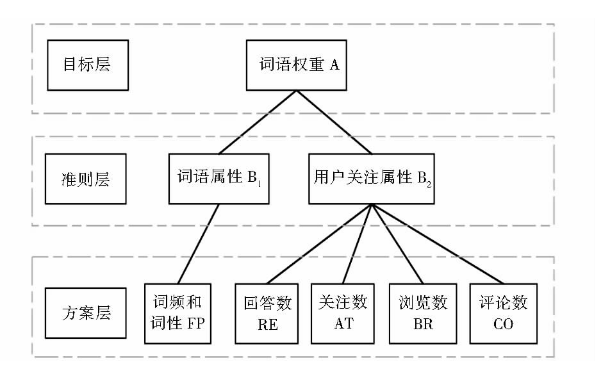



图 4 层次分析法结构模型

表 2 $B_2$ 判断矩阵

| | RE | AT | BR | CO | W | CR |
|---|---|---|---|---|---|---|
| RE | 1 | 1/3 | 3 | 1 | 0.203 | |
| AT | 3 | 1 | 5 | 3 | 0.526 | 0.001 |
| BR | 1/3 | 1/5 | 1 | 1/2 | 0.087 | 2 <0.1 |
| CO | 1 | 1/3 | 2 | 1 | 0.184 | |

3.4.3 层次单排序及一致性检验 引入随机一致性指标 RI 的 1～9 阶判断矩阵取值[31－32],见表 3。根据上述判断矩阵,使用方根法计算各因素的权重,结果如表 1、表 2 所示。由于 2 阶判断矩阵总是完全一致的,省去对矩阵 A 的一致性检验。经计算,矩阵 $B_2$ 的最大特征值 $\lambda_{max}=4.033\ 6$,一致性指标 $CI=(\lambda_{max}-n)/(n-1)=0.011\ 2$,查表 3 得知 $RI=0.96$,故 $CR=CI/RI=0.001\ 2<0.1$。一般地,$CR$ 值小于 0.1 就认为该矩阵具有满意一致性,因此矩阵 $B_2$ 通过检验。

表 3 n 阶判断矩阵的 RI 值

| 矩阵阶数 n | 1～2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| RI | 0 | 0.58 | 0.96 | 1.12 | 1.24 | 1.32 | 1.41 | 1.45 |

3.4.4 层次总排序及综合一致性检验 根据同一层次单排序的结果,能够计算各层要素相对于目标层的总权重。因此,从上到下将各层权重汇总并进行一致化处理得到各个指标的权重系数如表 4 所示:

表 4 层次总排序

| | $B_1$ | $B_2$ | 总排序权值 |
|---|---|---|---|
| | 0.5 | 0.5 | |
| FP | 1 | | 0.5 |
| RE | | 0.203 | 0.102 |
| AT | | 0.526 | 0.263 |
| BR | | 0.087 | 0.044 |
| CO | | 0.184 | 0.092 |
| CI | 0 | 0.011 2 | |
| RI | 0 | 0.96 | |

层次总排序的一致性比率 CR＝0.001 2＜0.1,说

明其具有满意一致性。因此,见公式 2,词语在单个文本中的权重计算中各个指标的系数 $\partial_1$、$\partial_2$、$\partial_3$、$\partial_4$、$\partial_5$ 的值分别为 0.5、0.102、0.263、0.044、0.092。

## 4 实验及分析

### 4.1 实验方法

知乎是国内社会化问答社区的代表[1]。在社会化问答社区中,用户的回答是围绕具体的提问展开的,提问内容能明确概括该问题下的回答内容。因此,本文根据知乎"汽车设计"话题下帖子的综合排序取前 1 000条,用八爪鱼采集器提取每条帖子的提问内容、问题标签、问题补充等文本数据以及回答数量、被关注数量、评论数量、浏览数量等数字数据,并将每条帖子的数据作为一个文本存放,去除重后得到文本 848 条,共 94 306 字。使用 HanLP 工具包进行分词、去停用词和词性标注等处理。同时,为进一步提高文本处理效果,从"汽车之家""太平洋汽车网"等平台收集"概念车""轮毂"等 41 645 个汽车相关词汇对 HanLP 词典进行了扩充。

根据文献[8]社会化问答社区话题识别与分析中研究方法的描述,借鉴黄鲁成等学者提出的关键词提取方法对本文的实验数据进行关键词提取,并把该方法表示为 COM。将 MW-KEM 的关键词提取性能与 COM、传统 TF-IDF 对比。通过控制文本数量、提取的关键词数量来设置两类实验,考察 3 种方法在不同条件下的关键词提取效果。第一类实验随机选取 N 条文本作为一个文本集,分别使用 3 种方法提取词语权重排序的前 N/3 个词作为该文本集的关键词,改变 N 值进行多次实验,考察 3 种方法在文本量变化时的提取能力;第二类实验以一个文本集为处理对象,分别使用 3 种方法提取规定数量的词作为关键词,改变提取的关键词数量进行多次实验,考察 3 种方法在关键词数量变化时的关键词提取能力。

### 4.2 结果分析

由于文本关键词提取方法的性能没有客观的评价指标,结合知乎中的问题标签进行人工标注,得到关键词参照集合。把机器方法提取的关键词和人工标注的关键词进行对比,使用准确率(Precision)、召回率(Recall)和 F 值(F-measure)对实验结果进行评价:

$$P = \frac{A \cap B}{A} \quad (8)$$

$$R = \frac{A \cap B}{B} \quad (9)$$

$$F = \frac{2PR}{P + R} \quad (10)$$

其中,A 指机器方法提取的关键词,B 指人工标注的关键词。

为对比 3 种方法应用于不同数量文本的效果,给 N 取不同值进行实验,得到实验结果如表 5 所示:

表 5 不同文本数量下 3 种方法性能对比

| N | 方法 | P 值(%) | R 值(%) | F 值(%) |
|---|---|---|---|---|
| 100 | MW-KEM | 72.7 | 49.0 | 58.5 |
| | COM | 60.6 | 40.8 | 48.8 |
| | TF-IDF | 63.6 | 42.9 | 51.2 |
| 200 | MW-KEM | 64.2 | 49.4 | 55.8 |
| | COM | 59.7 | 46.0 | 51.9 |
| | TF-IDF | 58.2 | 44.8 | 50.6 |
| 300 | MW-KEM | 63.0 | 53.4 | 57.8 |
| | COM | 53.0 | 44.9 | 48.6 |
| | TF-IDF | 54.0 | 45.8 | 49.5 |
| 400 | MW-KEM | 63.9 | 59.4 | 61.6 |
| | COM | 53.4 | 49.7 | 51.4 |
| | TF-IDF | 51.1 | 47.6 | 49.3 |
| 500 | MW-KEM | 63.9 | 60.9 | 62.4 |
| | COM | 51.2 | 48.9 | 50.0 |
| | TF-IDF | 50.0 | 47.7 | 48.8 |

结果表明,MW-KEM 的准确率、召回率、F 值均大于两种对比方法,说明其提取关键词的能力较强。绘制 3 种方法的 F 值随文本数量变化的趋势,如图 5 所示:

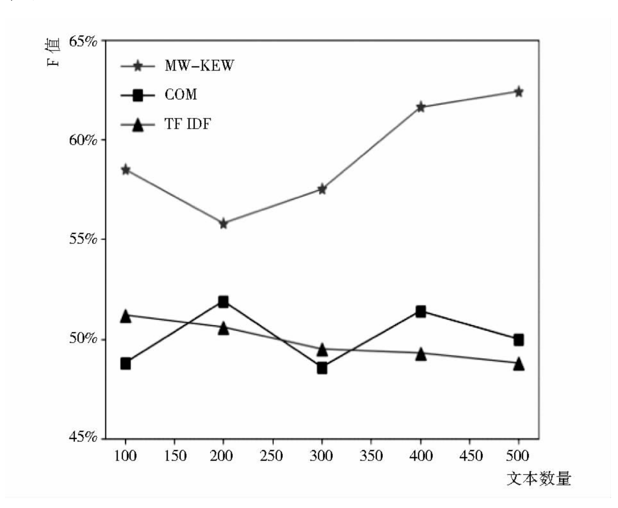



图 5 3 种方法 F 值随文本数量变化

从图 5 中可以看出:在处理社会化问答社区文本时,传统 TF-IDF 随着文本量的增加提取关键词的能力变弱;COM 与传统 TF-IDF 关键词提取能力相当;而 MW-KEM 不仅 F 值明显大于另两种方法,而且随文本

量的增加呈现增长趋势，反映了 MW-KEM 的良好性能。

为对比 3 种方法在同一文本集中提取不同数量关键词的能力，分别随机选取 200、500 条文本作为文本集，每个文本集人工标注 90 个词语作为关键词参照集合。分别采用 3 种方法提取 10、20，…，90 个关键词进行实验分析，部分结果见表 6。绘制文本量 200、500 下 3 种方法的 F 值随提取关键词数量变化的趋势，见图 6、图 7。实验表明，在关键词提取数量较小的时候，3 种方法的关键词提取能力相当：准确率高，但召回率较低，提取关键词的综合能力较差。随着关键词提取数量的增加，准确率降低，召回率上升，F 值呈现增涨趋势。当关键词提取数量大于 20 时，MW-KEM 的 F 值始终高于传统 TF-IDF 和 COM 方法，说明 MW-KEM 方法有更强的关键词提取能力。此外，文本量 200 条件下，MW-KEM 在关键词提取数量为 90 时 F 值为 54.2%，而在文本量 500 条件下，关键词提取数量为 90 时 F 值为 71.2%，说明了数据量的大小对提取效果有一定影响，也进一步验证了不同文本集下，MW-KEM 随文本量的增加关键词提取能力有所增强的趋势。

表 6　不同关键词个数下 3 种方法性能对比

| 关键词数 | 方法 | 文本量 200 | | | 文本量 500 | | |
|---|---|---|---|---|---|---|---|
| | | P 值(%) | R 值(%) | F 值(%) | P 值(%) | R 值(%) | F 值(%) |
| 10 | MW-KEM | 90.0 | 10.3 | 18.6 | 70.0 | 8.0 | 14.4 |
| | COM | 90.0 | 10.3 | 18.6 | 80.0 | 9.2 | 16.5 |
| | TF-IDF | 100.0 | 11.5 | 20.6 | 70.0 | 8.0 | 14.4 |
| 30 | MW-KEM | 76.7 | 26.4 | 39.3 | 80.0 | 27.6 | 41.0 |
| | COM | 70.0 | 24.1 | 35.9 | 76.7 | 26.4 | 39.3 |
| | TF-IDF | 66.7 | 23.0 | 34.2 | 73.3 | 25.3 | 37.6 |
| 50 | MW-KEM | 68.0 | 39.1 | 49.6 | 78.0 | 44.8 | 56.9 |
| | COM | 62.0 | 35.6 | 45.3 | 66.0 | 37.9 | 48.2 |
| | TF-IDF | 60.0 | 34.5 | 43.8 | 68.0 | 39.1 | 49.6 |
| 70 | MW-KEM | 62.9 | 50.6 | 56.1 | 74.3 | 59.8 | 66.2 |
| | COM | 60.0 | 48.3 | 53.5 | 68.6 | 55.2 | 61.1 |
| | TF-IDF | 57.1 | 46.0 | 51.0 | 58.6 | 47.1 | 52.2 |
| 90 | MW-KEM | 53.3 | 55.2 | 54.2 | 70.0 | 72.4 | 71.2 |
| | COM | 50.0 | 51.7 | 50.8 | 57.8 | 59.8 | 58.8 |
| | TF-IDF | 45.6 | 47.1 | 46.3 | 47.8 | 49.4 | 48.6 |

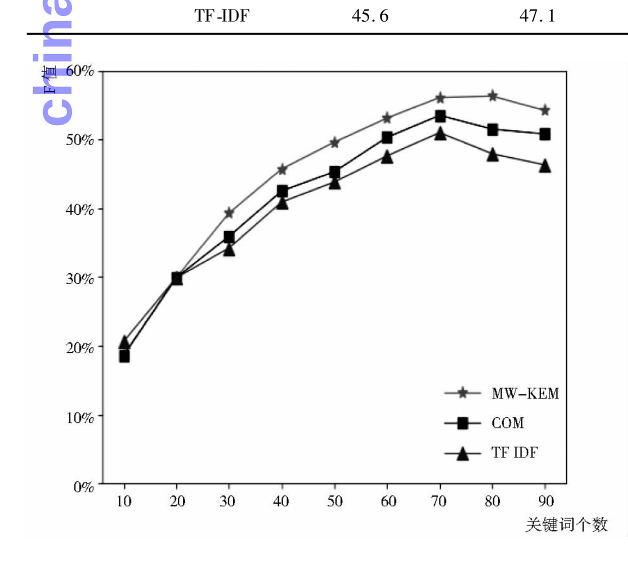



图 6　文本量 200 时 F 值随关键词数变化

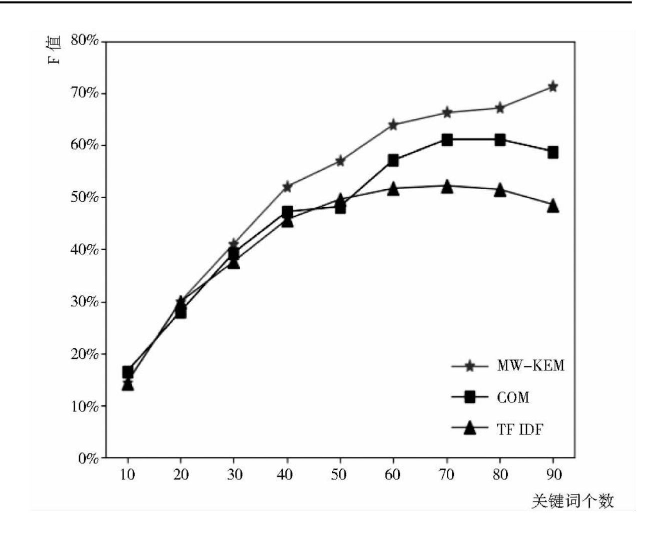



图 7　文本量 500 时 F 值随关键词数变化

通过不同文本量和不同关键词提取数量下的实验分析，能够验证 MW-KEM 方法在传统 TF-IDF 中引入调节函数、词性属性并融合用户关注属性，能有效提高社会化问答社区文本关键词提取的效率。同时，具备基于统计的关键词提取方法的优势：不需要大量的语料库和参数训练，简单便捷，又不依赖于语义背景知识

库,能有效保证关键词提取的客观性。

## 5 结语

本文在基于统计的关键词提取方法基础上,综合考虑社会化问答社区中影响词语权重的属性,提出适用于处理社会化问答社区文本的多属性加权关键词提取方法。该方法通过线性加权融合了词频、词性和用户浏览数、评论数等用户关注属性,并引入基于幂函数的调节函数的对传统 TF-IDF 方法进行改进来量化词语的词频和词性属性,引入 TF-PDF 方法量化用户关注属性。经验证,该方法能有效地提取社会化问答社区文本的关键词。

本文的研究方法也存在些不足:实验所用数据均为知乎"汽车设计"板块,数据来源单一,且较小数据集下关键词提取性能仍可继续提升。后续研究将在此基础上进一步提高关键词提取效率,同时,注重方法的实际应用,将社会化问答社区中提取的用户生成内容关键词与用户创新相结合,对用户创新的热点知识、核心知识和知识领域做分析识别。



**参考文献:**

[ 1 ] 陈娟,邓胜利. 社会化问答平台用户体验影响因素实证分析——以知乎为例[J]. 图书情报工作,2015, 59(24):102 – 108.
[ 2 ] 袁红,赵娟娟. 问答社区中用户与资源互动研究 [J]. 图书情报工作,2014, 58(18):102 – 109.
[ 3 ] WITTEN I H, PAYNTER G W, FRANK E, et al. KEA: Practical automatic keyphrase extraction [C]//Proceedings of the fourth ACM conference on Digital libraries. New York: ACM, 1999:254 – 255.
[ 4 ] HORITA K, KIMURA F, MAEDA A. Automatic keyword extraction for wikification of east asian language documents[J]. International journal of computer theory and engineering, 2016, 8(1):32 – 35.
[ 5 ] 方俊,郭雷,王晓东. 基于语义的关键词提取方法[J]. 计算机科学,2008,35(6):148 – 151.
[ 6 ] 费洪晓,康松林,朱小娟,等. 基于词频统计的中文分词的研究[J]. 计算机工程与应用,2005,41(7):67 – 68.
[ 7 ] 王立霞,淮晓永. 基于语义的中文文本关键词提取方法[J]. 计算机工程,2012,38(1):1 – 4.
[ 8 ] 黄鲁成,蒋林杉,苗红,等. 基于网络问答社区的话题识别与分析——以知乎"老年人"话题为例[J]. 图书情报工作, 2016,60(5):93 – 100.
[ 9 ] 陈娟,高杉,邓胜利. 社会化问答用户特征识别与行为动机分析——以"知乎"为例[J]. 情报科学,2017(5):69 – 74.
[10] 傅柱,王曰芬,陈必坤. 国内外知识流研究热点:基于词频的统计分析[J]. 图书馆学研究,2016(14):2 – 12.
[11] 罗燕,赵书良,李晓超,等. 基于词频统计的文本关键词提取方法[J]. 计算机应用,2016,36(3):718 – 725.
[12] 陈伟鹤,刘云. 基于词或词组长度和频数的短中文文本关键词提取方法[J]. 计算机科学,2016, 43(12):50 – 57.
[13] 张建娥. 基于多特征融合的中文文本关键词提取方法[J]. 情报理论与实践,2013,36(10):105 – 108.
[14] 张建娥. 基于 TFIDF 和词语关联度的中文关键词提取方法[J]. 情报科学, 2012(10):110 – 112,123.
[15] 张瑾. 基于改进 TF-IDF 方法的情报关键词提取方法[J]. 情报杂志, 2014(4):153 – 155.
[16] 廖晓,李志宏,席运江. 基于加权知识网络的企业社区用户创新知识建模及分析方法[J]. 系统工程理论与实践,2016,36(1):94 – 105.
[17] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5): 513 – 523.
[18] PAIK J H. A novel TF-IDF weighting scheme for effective ranking [C]//Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2013:343 – 352.
[19] 施聪莺,徐朝军,杨晓江. TFIDF 方法研究综述[J]. 计算机应用,2009,29(s1):167 – 170.
[20] 罗繁明,杨海深. 大数据时代基于统计特征的情报关键词提取方法[J]. 情报资料工作,2013,34(3):19 – 20.
[21] 钱爱兵,江岚. 基于改进 TF-IDF 的中文网页关键词抽取——以新闻网页为例[J]. 情报理论与实践,2008, 31(6):147 – 152.
[22] 张保富,施化吉,马素琴. 基于 TFIDF 文本特征加权方法的改进研究[J]. 计算机应用与软件,2011,28(2):17 – 20.
[23] 袁津生,毛新武. 基于组合特征的中文新闻网页关键词提取方法[J]. 计算机工程与应用,2014, 50(19):222 – 226.
[24] 蒋昌金,彭宏,陈建超,等. 基于主题词权重和句子特征的自动文摘[J]. 华南理工大学学报(自然科学版), 2010,38(7):50 – 55.
[25] 李湘东,巴志超,黄莉. 一种基于加权 LDA 模型和多粒度的文本特征选择方法[J]. 现代图书情报技术,2015,31(5):42 – 49.
[26] 路永和,王鸿滨. 文本分类中受词性影响的特征权重计算方法[J]. 现代图书情报技术,2015,31(4):18 – 25.
[27] 周鹏,蔡淑琴,石双元,等. 基于关键词抽取的微博舆情事件内容聚合[J]. 情报杂志,2014(1):91 – 96.
[28] YE H M, CHENG W, DAI G Z. Design and implementation of on-line hot topic discovery model[J]. Wuhan University journal of natural sciences, 2006, 11(1):21 – 26.
[29] SAATY T L. Modeling unstructured decision problems-the theory of analytical hierarchies[J]. Mathematics and computers in simulation, 1978, 20(3):147 – 158.
[30] 刘开第,庞彦军,周少玲,等. 多准则排序中的路径问题及层次分析法推广[J]. 系统工程理论与实践, 2015,35(4):973 – 983.

[31] 邓爱东.多层次模糊综合评价法在图书馆危机管理中的应用[J].现代情报,2008,28(6):117－119.
[32] 李亚平,焦建玲.网上交易流程效率评价[J].合肥工业大学学报:自然科学版,2009,32(8):1204－1207.

**作者贡献说明:**

余本功:指导研究思路的设计,对研究方法、研究过程给予修正,提出修改意见;

李婷:设计研究思路,数据采集及实验分析,撰写并修改论文;

杨颖:指导实验研究,对论文修改提供思路和建议。

## Keywords Extraction Method for the Social Q&A Community Based on Multi-attributes Weighted

Yu Bengong[1,2]　Li Ting[1]　Yang Ying[1]
[1] School of Management, Hefei University of Technology, Hefei 230009
[2] Key Laboratory of Process Optimization & Intelligent Decision-making, Ministry of Education, Hefei University of Technology, Hefei 230009

**Abstract:** [**Purpose/significance**] Existing methods of extracting keywords can't be applied to the social Q&A community effectively, because they are not suitable for the characteristics of the social Q&A community which embodies short texts, colloquial contents and sparse data. They rarely think about the impact of users' attention on words. In view of the aforementioned problem, this paper presents a novel keywords extraction method based on multi-attributes weighted for the social Q&A community. [**Method/process**] This method improved the traditional TF-IDF algorithm by introducing the tuning function and the part of speech. Besides, it calculated the weight of words based on a linear weighting formula, which fused four attributes of user focus by dealing with numbers of users' answer, attention, browse, and comments. [**Result/conclusion**] Experiments show that this method can extract keywords from the social Q&A community more effectively.

**Keywords:** social Q&A community　keyword extraction　TF-IDF　multi-attributes weighted